Journal of Educational and Behavioral Statistics
Spring 2004, Vol. 29, No. 1, pp. 117–120

The Real World is More Complicated than We Would Like

Mark D. Reckase
Michigan State University

It is understandable that parents, policy makers, educators, etc. want to know how schools are functioning. Extensive resources are expended on the educational enterprise and it is only reasonable that the impact of those resources be determined. However, determining the amount of change in students' skills and knowledge is not easy. Further, there is a desire to use relatively simple models to represent and report the results, but the reality is not as simple as we would like and using the simple models for the purposes of making reports understandable may lead to misleading interpretations. To illustrate these points, a simple model of growth is provided and its use is discussed to show why it can lead to misleading results. Then, the more complex situation of growth measurement using educational assessments is considered. Finally, the implications for value-added analyses will be addressed with some suggestions about how using over simplified models may lead to misleading results.

A Simple Conception of Growth

A simple conception of the measurement of growth is marking the height of a child on a wall to show the increase of height with age. In the old days, this would be done by making a pencil mark directly on the wall or by making notches in a door frame. Now there are paper charts taped to the wall with the foot and inch scale indicated on the chart. From the marks on the wall, or chart, it is easy to see the amount of growth and when the growth occurred if dates are attached to the specific marks.

Many educators and policy makers would like to have a similar chart for students' growth in academic subjects. As students progress in learning reading, mathematics, science, etc. a chart could be used for each student with marks showing the amount of growth at specified points in time. To support such uses, test developers have produced vertical scales for academic subject matter areas that cover multiple grades. Although there are warnings against over interpreting these scales, the existence of these scales seems to indicate the simple growth chart for height can be used as a model for reporting academic learning.

But the height chart does not work as well as a measure of growth as it at first seems. As children grow, they grow in more ways than height. They increase in weight and girth. The lengths of arms and legs and head sizes change as well. In fact, there is usually a point when children stop increasing in height. After that point in

time, they usually continue to increase in weight and girth. For me personally, I have stopped increasing in height and weight, but somehow the circumference at my waist continues to increase. To accurately model growth, we need to think about change in a multidimensional way rather than using a single unidimensional scale.

Although this will seem unreasonable here, it is possible to create a growth scale for use with children that shifts from height to weight as the increase in height slows to a stop. Such a scale is typically not considered because the numerical scales for height and weight are so different. But suppose that the scale for weight were converted to a new scale with a range from 0 to 84 and the same mean and standard deviation as height measures in inches for students at, say, 14 years of age. Then, after age 14, weight could be used to continue the growth trend started with 14 years of height data.

For young ages, changes in height are highly correlated to changes in weight so it might be reasonable to use one as a proxy for the other. For example, grade-4 weight might be used to predict grade-4 height and the difference between the predicted height and the grade-5 height might be a reasonable indicator of growth in height. That process would not work to assess growth for 20-year-olds, because the change in weight is no longer highly correlated with change in height. The relationship between the variables is not the same for the different age groups.

Growth in Academic Domains

Of course, such growth scales and growth predictors do not seem reasonable, but they do mirror what is done with vertical scales for academic areas. For mathematics, for example, tests at the 3rd-grade level measure predominantly arithmetic skills. By 8th grade, the test shifts to problem solving, pre-algebra, and algebra skills. Yet, the way that results are reported on the vertical scales seem to imply that the tests are measuring the same thing.

The real situation is actually more complicated then this cross grade example. Within the tests at a grade, the difficulty items may be measuring different combinations of skills than the easy items. For a 10th-grade mathematics test, the hard items tend to be from coordinate geometry with a heavy emphasis on manipulating abstract mathematical content. The easier items tended to deal more with arithmetic problems solving and computation. As a result, differences in scores at the bottom end of the score scale indicated differences in arithmetic problem solving skills while differences at the top end of the score scale indicated differences in skills for manipulating abstract mathematical concepts (Miller & Hirsch, 1992). This is not a serious problem for this test, because care is taken to make all test forms provide this same pattern of relationships. However, it is unlikely that vertically-scaled, grade-level tests have been analyzed to discover the multivariate structure and the relationship of that structure to item difficulty, or that the creation of multiple forms takes these relationships into account.

The implication of these types of findings is that the simple model of the height chart does not apply to academic testing. Rather, growth in student performance may take a circuitous path through many domains of test content. The tests that are

used may or may not reflect the actual path of change in knowledge and skills, but they are more complex than a single linear continuum. These complexities need to be taken into account in growth modeling. Perhaps the change in student performance should be measured along a curve that goes through a number of content dimensions instead of along a single linear continuum.

Many years ago while working in multidimensional item response theory, Wang (1986) showed that the result of applying a unidimensional item response theory model to data from test items that vary in the dimensions to which they are sensitive is a linear composite of the dimensions in the data. That is, complex relationships in the data are projected onto a line and the particular line is a weighted composite of the dimensions in the data. If the dimensions of sensitivity of the items change, such as an emphasis on geometry instead of algebra, the characteristics of the weighted composite will change. Thus, conceptually, the score scales from tests are linear, but the lines change in orientation as the tests are designed for successively higher grade levels. With enough short line segments a good approximation can be made to a curve. Of course, projecting the complex data onto a line results in the loss of information and when that linear scale is extended over many grade levels, the loss of information might be extensive.

The Influence of Test Structure on Value-Added Assessment

The value-added assessment framework seems to be built on either the growth chart model or a residualized gain model much like predicting height from weight. All of the articles in this issue of *JEBS* discuss growth in performance and most use a simple difference score as an indicator of growth. If the model just presented is an accurate depiction of the ways that tests really function, then the meaning of score differences needs to be questioned. The example of the mathematics tests suggests that the difference score may be the difference between pre-algebra skills and coordinate geometry skills. These differences may be very difficult to interpret. Further, if a teacher emphasizes pre-algebra in a course, but not coordinate geometry, there may be improvements in performance that are not shown on the test because of its shift in emphasis. This shift in test content likely occurs both within tests and across grade-level test forms. Maximum change in test scores will occur if the pattern of instruction matches the shift in the construct assessed by the test. A mismatch between instruction and the assessment will result in an underestimate in students' change in performance. The equivalent from our initial example is indicating that there is no growth because height is constant even though there is a change in weight.

In one sense, this is not a problem with the methods used for analysis because the numerical values used in the procedures generally meet the distributional assumptions of the procedures. The problem comes in the interpretation of the results. Growth is defined as the difference between two numbers. Is it reasonable to compute such differences when the numbers mean different things? Rather than ignore the characteristics of the sources for the scores, statisticians need to consider the meaningfulness of the numerical scales and how that meaning changes

over grade levels. Ultimately, nonlinear multivariate models may be needed to track the changes in educational performance. More care is probably also needed to align the tests to the actual path of growth in the academic area for students. The sophisticated statistical procedures described in these articles may be giving a glossy finish to misleading assessment results. Before putting a lot of confidence in the results of these analyses, the functioning of the assessments needs to be investigated in great detail.

References

Miller, T. R., & Hirsch, T. M. (1992). Cluster analysis of angular data in applications of multidimensional item response theory. *Applied Measurement in Education, 5*(3), 193–212.

Wang, M. (1986, April). Fitting a unidimensional model to multidimensional item response data. Paper presented at the Office of Naval Research contractors meeting.

Author

MARK D. RECKASE is Professor, Michigan State University, 461 Erickson Hall East Lansing MI 48824-1034; reckase@msu.edu. His areas of specialization are measurement and quantitative methods.